ENGINEERING LINGUISTICS

Valentin CIJACOVSCHI,
PhD habilitatus, University Professor, ULIM,
Vladimir PROCOPCIUC,
PhD student, TUM, R.Moldova

Engineering linguistics as a compound part of the general linguistics has a special place due to the fact that it is meant to solve specific issues, using programming automaton on one hand, the main problem being the automatic processing of the linguistic text, and on the other hand using the theory of linguistic synergy.

It can be understood that the person who organizes and carries the above mentioned process must be able to combine two specialties: that of linguist and that of programmer.

Some general requirements for these specialists were partially stated by *International Center for Applied Business Intelligence* (ICAPI: recrutement@icapi.com) in their job offer for engineers in natural language machine processing (NLMP). This centre has also the following requirements:

- Capability to work in a team of linguists, engineers and developers;

- Clearly understanding of linguistic means;
- Imagination and realization of the innovative implementations of using the technologies for natural language processing.

Thus it becoming clear, accumulating by one person the specialty mentioned above such a problem could be successfully solved.

In order to correspond to requirements stated by ICAPI, the future specialist, engineer-linguist should learn a set of knowledge addressed to two specialties: in specialization in domain of theory and practice of engineer linguistic and in the domain of theory and practice of programming.

Specialization in domain of theory and practice of engineer linguistic

- diffuse nature of linguistic objects;
- linguistic models;
- linguistic sign and sign in artificial language;
- verbal-cognitive communicative process;
- speech and language synergy;
- linguistic automatons;
- artificial intelligence and its linguistic aspects;
- Databases and knowledgebase.

Specialization in theory and practice of programming

The base knowledge for this specialty consists from:

- programming;
- databases and algorithms;
- linguistic mathematics;
- object oriented programming;
- logical programming;
- formal languages;
- databases and knowledgebase;
- functional programming;
- engineer linguistics – as source of ideas for programming;
- Systems of artificial intelligence.

For attending this specialties students from foreign language, informatics and programming departments are invited.

Education of such specialists is divided into two phases:

- Beginner's course phase (training phase) from corresponding departments and
- Master course phase, when they'd obtain scientific degree of linguist-engineer / engineer-linguist.

Theoretical basis

Is the concept of *linguistic synergetics*. *Synergetics* or *synergism.* (The term is derived from the *Greek* *syn-ergos*, *συνεργός* meaning working together.) was related to the phenomenon in which two or more influences (two or more agents), activating together, obtain an effect greater than that can be predicted knowing just the effects of individual agents. From the beginning it was a scientific term.

The synergy may also mean:

- A mutual advantageous link, the sum of which is greater than the sum of their component parts;
- A dynamic state in which combined action is favored over the difference of individual component actions.;
- Behavior of whole systems unpredicted by the behavior of their parts taken separately. More accurately known as emergent behavior.

Strictly speaking it's about a phenomenon which is obtained as a result of the actions of some different factors. The last taken separately doesn't give the mentioned phenomenon. Theory of dynamic systems presents the method or mathematic apparatus used in synergy.

As an example of a efficient collaboration of a linguist and a programmer, we propose the description of the experiment of machine translation "*English business letter machine translation without post-editing*" from English into German.

We ascertain the fact that the existent systems of machine translation such as Systran, Global Link, and others don't achieve a fine result in translation that would not need any man intervention, and namely:

- an automatic translation without post-editing isn't obtained;
- the translation is based on improving the existing morphologic-syntactical data without taking into account the possibilities of a semantic analysis based on the synergetic laws of the natural language;
- the translation is realized by programs without a corresponding understanding of the translated text's essence.

To obtain a method of machine translation that wouldn't need any post-editing, we are relying on the possibilities to implement theoretical and practical ideas of engineer linguistics, the last allow an adequate interpretation of some concepts such as

- artificial and natural intelligence,
- linguistic synergy and methods of analysis,
- transfer and synthesis of specialized texts exposed to machine translation

The above-mentioned process is formed up by three stages; first one consists of the following procedures:

- Making up the vocabulary (English and German);

- Duality elimination of some parts of speech;
- Segmentation of the English sentence;
- Schematization of the English sentence;
- Disambiguation of the parts of speech.

The second stage consists of:

- Recoding of the German parts of speech;
- Words order fixation for the German sentence;
- Determination of the *Exterior and Interior Valence Determinative Ties* (EVDT, IVDT)

The third stage consists of the final translation itself. Translation procedures of each section are based on linguistic algorithms transferred into mathematical algorithms and finally programmed.

As a result of all the operations described above, the program becomes able to determine the exact German form suitable to English one and at the same time the filtration of German equivalents is performed, taking into consideration the polysemous character of corresponding English word forms. There are applied different methods when choosing the right equivalent and among them is theory and practice of terms' disambiguation.

The apparition in the late 40's of the computer capable of solving not only calculation tasks but also performing logical operations made possible its utilization in translating. As a result, already in the early 60's the linguistic-engineering scientific group formed up which united linguists and mathematicians-programmers and was pursuing the goal of creating computerized models for reproduction of different linguistic phenomena and some aspects of the verbal-cognitive human activity (VCHA). As the main impediment for realization of this goal was considered the behavior of the natural languages (NL) and programming languages. Disregarding this fact has led to the failure of the system TA EUROTRA in the 70-80's. Unlimited concentration of different language and speech details could not favor the automatic processing of the text (*First International Conference on Language Resource and Evaluation, Spain, 1998*).

One of the latest and often used automatic translating systems is the American system SYSTRAN (*Linguistic Description of Systran „Luxemburg Commission of the European Communities, April 1993"*). It is based on the use of automat-statistical dictionaries of big volume and of simplified grammatical frames. Based on this system we can obtain translations of big volume but with a poor quality which doesn't satisfy users and needs post-editing.

Procedures description

Domain of application „computer solving of linguistic issues":

Segmentation- automatic segmentation of initial natural language (INL) sentence based on the system of codification of INL parts of speech, set of rules for duality elimination of interpreting and of table of indices;

Schematization- automatic assignation of sentence's segments already detected by the linguistic cybernetic automat (LCA) their syntactic-semantic functions simultaneously recording their graphical sequence order;

Disambiguation- (*segments' disambiguation*) automatic elimination of the polysemy peculiar to the majority of English parts of speech, based on the set of disambiguation rules;

Recoding - recoding of the *initial natural language* (INL) segment into *final natural language* (FNL) in case of lacking of coincidence between INL and FNL expression forms;

Ordering - word-order settlement in the FNL;

Filtration - choosing the FNL translation equivalent word suitable for INL;

EVDT (*Exterior Valence of Determinative Ties*) - setting and recording the general exterior valence;

IVDT (*Interior Valence of Determinative Ties*) -setting and recording the general interior valence;

The subject of study is a new procedure based on the utilization of the natural language synergetic mechanism, in this case of INL and FNL specialized in a certain domain, in order to ensure an automatic translation of the specialized INL texts into FNL without post-editing.

The machine translation based on utilization of the linguistic synergetic mechanism is characterized with the fact that the machine translation is carried out without the necessity of post-editing.

The proposed procedure for machine translation of specialized English/German (INL) into FNL is based on forming of the generating and perceiving operations of these texts in human mind, formalization of these operations in the form of algorithms and their operative programming for their posterior regeneration by the LCA.

The distinctive difference from the existent procedures of machine translation consists in the fact that the operations sequence for solving this issue allows automatic intimation of the functional manner of the linguistic synergetic mechanism created in an ideal manner by setting up the transition forms into a single definite and materialized form at the level of words, junction of words (segments) and sentence (junction of segments).

Claims:

1. Automatic segmentation of the English (INL) sentence based on the codification system of English parts of speech, rule set of duality elimination and table of indices;

2. schematization, thus automatic assignation of the segments of the sentence already detected by LCA, of their syntactic-semantic functions and simultaneously recording their graphic sequence order;
3. segment's disambiguation, thus automatic elimination of the polysemy peculiar to the majority if English parts of speech, based on the set of disambiguation rules;
4. recoding of the INL segment into FNL in case of lacking of coincidence between INL and FNL expression forms;
5. word-order settlement in the FNL;
6. filtration, thus choosing the FNL translation equivalent word suitable for INL;
7. fastening of the morphological and syntactical constant characteristics of the parts of speech of the analyzed FNL segment;
8. setting and recording Exterior Valence of Determinative Ties (EVDT);
9. setting and recording Interior Valence of Determinative Ties (IVDT);
10. final recording of the morphological and syntactical constant characteristics including inconstant ones;
11. automatic choosing the translation variant, thus recording the paradigms of the declination and conjugation of the needed word form from the segment where it is located;
12. final fixation of the translation result from INL into FNL.

Realization stages of text machine processing of INL for its translation into FNL

Introductory remark

In the following section is described the corresponding order of the machine processing of specialized text/sentence of INL, i.e. the sequence and content of the elements of this continuous process which ensures the transition to translation into FNL.

It includes: segmentation, preceded by elimination of interpretation duality of some parts of speech, schematization and disambiguation. All these operations assume the presence of codification of the English (INL) parts of speech, of rules set which regulates their realization as linguistic as well as programming.

Automatic segmentation of INL sentence

With notion of segmentation it is understood automating setting of the component parts of content of the analyzed sentence which takes part in making up its meaning. The segment can be represented as a part of speech taken apart (conjunction *while*) or as a junction of words (*your proposal*).

The automatic segments' determination is preceded by elaboration of the engineer reproductive models (ERM) which simulates the order of generating and perceiving of the analyzed sentence dictated by the linguistic

synergetic mechanism itself, NL and VCHA. These models are based on utilization of morphological and syntactic-semantic properties of the English (INL) parts of speech which manifest themselves differently in dependence on the role played in creating the meaning of the current sentence.

Rules set for elimination of the interpretation duality

In cases when an elimination of interpretation of one ambiguous part of speech is necessary before the operation of automatic segmentation the 3 letter of the code A will be replaced by R.

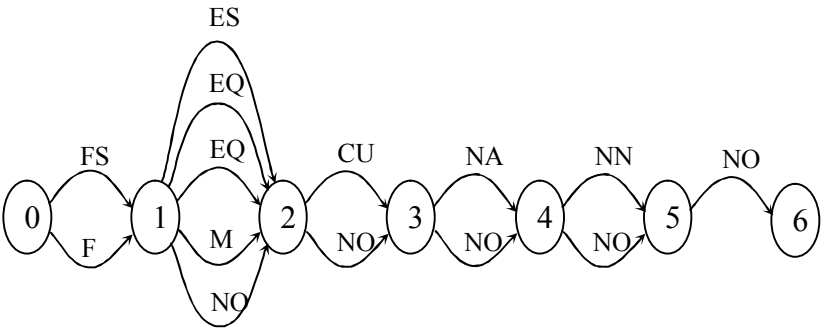
For example: the preposition *to* is coded with *PTR*, as soon as it can introduce an infinitive, or a prepositional segment. In first case *R* will be replaced by *O* (*PTO*), in the second case *PTR* will be replaced by *PSO* (simple unambiguous preposition).

Participial conjunction *considering that*, when detected by the program with its first member *considering* coded *VGR*, is to be processed in concordance to respective rule. The latest supposes the detection of the following part of speech *that* (*CTA*) in case of its detection allows the fixation of the conjunctional segment coded with *CPO* using two intervals (␣␣).

Tables of indices are a systematization of codes arranged in corresponding columns (1, 2, 3, ...) of the English (INL) parts of speech which take part in building up the semantic segments of the analyzed sentences and numbered from 0 till 9 according to number of the English (INL) coded parts of speech. In case of extending the including codes the table can be extended (for pronouns we introduce 4A, 4B, 4C). As an example we shall use the table of indices from bellow.

Table of indices no. 4B

l. no.	1	2	3	4	5	6	7	8		10
Parts of speech codes	FSO FAO	ESO EQO EQA MOA NOA	CUL NOA	NAO NOA	NNO NOA	NOA	-	-		



This table refers to all segments introduced by the possessive pronoun. It is coded as *FSO* when it is used as an adjectival pronoun (associative form) and as *FAO* when it is used as a noun (absolute form). Both codes take first column. In the next columns there are indicated the codes of English (INL) parts of speech which can participate in building up the segments introduced by *FSO* and *FAO* (adjective is introduced with *E*, numeral *with M*, noun with *N*, conjunction with *C*).

For example, the program begins with processing of the segment introduced by the possessive pronoun (*FSO*) (see table of indices no. 4B p. 3). Marking its presence in the first column of the table, the program records its presence in one interval (*your*␣) and pursues to processing of the next word - *proposal* – coded as *NOA*. In order to decide its membership to the segment it is necessary to address to the second column for confirmation of its presence. The answer being affirmative, the program records it in one interval and pursues to the processing of the net word from the sentence – *for* (*PSO*). Addressing to the table of indices is repeating, but this time to the third column. The absence of *PSO* in the column is interpreted by program as the beginning of the new segment which should be separated from the previous with two intervals (␣␣) - *your*␣*proposal*␣␣.

In the same way was fixed the end of the first segment of the analyzed sentence introduced by *we* (*FNO*), but this time addressing to the table of indices 4A.

We used 16 tables of indices, 5 for segments introduced by verb, 3 for segments introduced by pronouns and one for the each rest of the parts of speech, all these tables are projected in database with respective architecture for recording the data for graphs.

General remarks

1.The proposed automatic translation experiment of INL specialized texts into FNL is dedicated to simulation of process of morphologic-

syntactic-semantic analizing and synthesizing realized in human mind when generating and perceiving the specialized and standartized text in correspondence to architectural structure of the program through algorythmisation and programming.

2. Taking into account the complexity of this process, we have divided and subdivided it into linguistic units and subunits which make possible a successive solving of the entire problem when the preceding results allow solving the following issues.

3. It is one of the first research-work of this kind which addresses directly to the functioning of the mechanism of synergetic linguistics at the implementation of automatic processing of specialised text. As a result we have marked the existence of interior interdependence relations between linguistic and psychic phenomena which ensure the functioning of the above-mentioned mechanism. While our thoughts are produced in an ideal form and are properties of the substance organised in a particular mode our reasons, notions, conclusions and the natural and artificial languages are produced in a concrete and materialised form – grammar structures and vocabulary.

The mutual dependence between language as a phenomenon on one hand and materialized on the other hand based on automation of the specialized texts allowed affirmation of machine translation without post-editing.

Bibliography

1. Cijacovschi V., Procopciuc V. Lingvistica inginerească - component contemporan al lingvisticii (realităţi şi perspective). Chişinău: Centrul Ed. USM, 2005.

2. Tschizhakowski V., Schtserbina O., Popescu A. Bildung des Terminologischen System im Fachgebiet «Wirtschaft». Chişinău: ULIM, 2003.

3. Cijacovschi V., Procopciuc V. English business letter machine translation without post-editing. Chişinău: USM, 2005.

4. Tschizhakowski V., Schtscherbinina O., Popescu A. Bildung des terminologischen lexisch-sematischen Systems im Fachgebiet „Wirtschaft" (die deutsch-russische Variante)", Chisinau: Freie Internationale Universität Moldovas, 2002.

5. Пиотровский Р.Г. Инженерная лигвистика и теория языка Ленинград: Наука, 1979.